MAS223 Statistical Inference and Modelling

Dr Nic Freeman

August 19, 2016

Introduction

The course material consists of:

- These **lecture notes**.
- A booklet of **examples**, which accompany these notes. Examples are referred to in blue.
- A booklet of **exercises**, containing one set of exercises for each chapter. Typed **solutions** to the exercises will be provided online. Exercises are referred to in bold, e.g. **Q1.23**.
- A **formula sheet** which contains helpful formula and other information about a selection of standard distributions.

The full content of the course is covered in the typed notes. There is no need to take handwritten notes in lectures, although you may wish to annotate the typed notes and examples as you become familiar with them. Naturally, it is also important to work through the exercises.

In **Chapter 1** we develop the theory of (univariate) random variables, following on from first year courses. We focus on continuous random variables; discrete random variables were covered in MAS113. Then, in **Chapter 2** we build up a library of standard distributions. Our goal is to have a supply of useful distributions for future use, both for later chapters and for future probability and statistics courses.

In **Chapter 3** we examine transformations of univariate random variables, meaning that if X is a known random variable and g is a (non-random) function, we look to obtain information about $g(X)$. This allows us to record many useful relationships between the standard distributions of Chapter 2.

We move on to study multivariate random variables in **Chapter 4**, extending the univariate theory covered in Chapter 1. Again, we focus on continuous random variables, introducing ideas such as independence and conditional probability. We study transformations of multivariate random variables in **Chapter 5**, extending the univariate theory covered in Chapter 3.

In **Chapter 6** we study the multivariate normal distribution, which generalizes the normal distribution into $\mathbb{R}^d$. The importance of the multivariate normal distribution to stochastic modelling cannot be overstated; it is a popular tool in very many areas of statistics.

Chapter 7 moves away from probability theory and into statistical inference. We introduce the idea of likelihood and then focus on maximum likelihood, which is a method of choosing parameter values so as to fit stochastic models to data. Lastly, in **Chapter 8**, we look at some case studies (taken from the recent literature) in which the tools we have developed are used to draw conclusions from real world data.

Contents

Chapter 1

Univariate Distribution Theory

We start with some revision of material from first-year courses, in particular MAS113 Introduction to Probability and Statistics.

In probability and statistics we are usually interested in situations where there is some uncertainty about the outcome. We often refer to such situations as experiments. We identify a set S of possible outcomes, known as the **sample space**; one and only one of these possible outcomes will actually occur when the experiment is performed. If we repeat the experiment, the outcome may change.

Events are subsets of the sample space S. If $A \subseteq S$ is an event, then the 'true' outcome may or may not be a member of A; if it is, we say that A occurs. To every event we associate a probability, $\mathbb{P}[A]$, which we think of as the chance of the event A occurring.

Frequently, we are interested in a numerical measurement arising from an experiment, rather than the raw outcome - for example, we might count the number of heads in a sequence of coin tosses, rather than recording the exact sequence of heads and tails. In such situations we work with a **random variable** X, which is a function $X : S \to \mathbb{R}$. Then, X associates each element of the sample space to a real number, and we are interested in probabilities of the form $\mathbb{P}[X \in E]$, where E is a subset of $\mathbb{R}$. These probabilities form the **distribution** (or probability distribution) of the random variable.

We write R_X for the **range** of X; this is precisely the set of values that the random variable X may take.

Example 1: *Sample spaces and random variables*

The most important property of a random variable is its distribution.

Definition 1.1 *The **distribution function** of the random variable X is the function $F_X : \mathbb{R} \to [0, 1]$, given by*

$$F_X(x) = \mathbb{P}[X \leq x].$$

The function F_X is also sometimes referred to as the cumulative distribution function. When it is clear which random variable we mean, we will often drop the subscript and write $F = F_X$.

Most distributions which we encounter come from two special types; discrete random variables and continuous random variables.

1.1 Random variables

Definition 1.2 *If a random variable X is integer valued (or, more generally, takes values only in some finite or countable set), then we say X is a* ***discrete random variable****.*

In the discrete case, we write

$$p(x) = \mathbb{P}[X = x],$$

and we call p the probability function of X. The graph of F increases entirely by jump discontinuities, jumping upwards at each x for which $p(x) > 0$. The size of the jump at x will be $p(x) = F(x) - F(x-)$.

Example 2: *Discrete random variables.*

Many random variables (the normal distribution, for example) take values in $\mathbb{R}$, which is not countable. For these cases, we need a more sophisticated way of describing random variables.

Definition 1.3 *A function $f : \mathbb{R} \to \mathbb{R}$ is a* ***probability density function*** *if both*

1. *$f(x) \geq 0$ for all $x \in \mathbb{R}$,*
2. *$\int_{-\infty}^{\infty} f(x)\,dx = 1$.*

If a function $f(x)$ is a probability density function, then there is a random variable X such that the distribution function of X satisfies $F_X(x) = \int_{-\infty}^{x} f(u)\,du$. Proving this fact (i.e. the existence of X) requires some analysis and is outside the scope of our course. However, it allows us to make the following definition.

Definition 1.4 *If we can write the distribution of a random variable X in the form*

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx \tag{1.1}$$

where $f_X(x)$ is a probability density function, then we say X is a ***continuous random variable****. We call f_X the probability density function of X.*

If we know the distribution function of a random variable, then we can find its probability density function using

$$\frac{d}{dx}F(x) = f(x).$$

Conversely, if we are given f, then we can use (1.1) to find F.

In the continuous case, probabilities may be found by integrating the p.d.f. over an appropriate range. For example,

$$\mathbb{P}[x < X < y] = F(y) - F(x) = \int_{x}^{y} f(t)\,dt. \tag{1.2}$$

In the continuous case, we have $\mathbb{P}[X = x] = \int_{x}^{x} f(u)\ du = 0$ for all x (but in the discrete case we can have $\mathbb{P}[X = x] > 0$).

We will often encounter probability density functions $f(x)$ that are defined by different formulae for different ranges of x. In these cases, we can still calculate probabilities using (1.2), but to calculate the integral we must first the split it up into the different intervals for each formula.

Example 3: *Continuous random variables and probability density functions.*

1.2 Distribution functions

From the distribution function of X, we can calculate the probability of more complicated events. For example if $x < y$ then

$$\begin{aligned} \mathbb{P}[x < X \leq y] &= \mathbb{P}[X \leq y] - \mathbb{P}[X \leq x] \\ &= F_X(y) - F_X(x). \end{aligned}$$

If we know the distribution of a random variable, then we can (in principle) use it calculate any probability associated to that random variable.

For a general function $F : \mathbb{R} \to [0, 1]$, we say that F is a **distribution function** if it has has the following properties.

1. $0 \leq F(x) \leq 1$ with $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1$.
2. $F(x)$ is non-decreasing in x; that is, if $x < y$ then $F(x) \leq F(y)$.
3. F is right-continuous and has left limits.

It can be shown that, for any random variable X, its distribution function satisfies 1-3. Conversely, if we have a function F satisfying properties 1-3, it is also true that there exists a random variable X with distribution function F. Proving these facts requires some analysis, and is outside of the scope of this course.

A probability density function f must be non-negative because F cannot decrease, but note that a probability density function f is not itself a distribution function. It is possible (and common) for f to be greater than 1 for some values of x.

Example 4: *Properties of distribution functions*

1.3 Means, variances and moments

In the discrete case, we define the **mean** (or **expectation** or **expected value**) of the random variable X to be

$$\mathbb{E}[X] = \sum_{x\in R_X} xp(x). \tag{1.3}$$

Here, R_X denotes the range of values that the random variable X can take. Similarly, in the continuous case, we define

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\, dx. \tag{1.4}$$

Comparing these two formulas, we might think of $\int \ldots dx$ as a 'continuous version' of $\sum_x \ldots$, and we might think of the p.d.f. $f(x)$ as a continuous equivalent of p.f. $p(x)$. That is, we think of $f(x)$ as a measure of how likely X is to be 'nearly' equal to x.

More generally, if $g(X)$ is a function of X then

$$\mathbb{E}[g(X)] = \sum_{x \in R_X} g(x)p(x) \qquad \text{(discrete case)},$$
$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \quad \text{(continuous case)}. \tag{1.5}$$

We often write $\mu = \mu_X = \mathbb{E}[X]$ for the mean. Note that, setting $g(x) = x$, we recover the formulae for $\mathbb{E}[X]$. Taking $g(x) = x^r$, where $r \in \mathbb{N}$, we obtain a formula for the **$\mathbf{r^{th}}$ moment**, $\mathbb{E}[X^r]$.

With special choices of g, we can extract important information about the random variable X. One especially useful quantity is the **variance**

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}[X^2] - \mu^2.$$

We often write $\sigma^2 = \sigma_X^2 = \mathrm{Var}(X)$. The positive square root $\sigma = \sqrt{\mathrm{Var}(X)}$, is known as the **standard deviation**.

Example 5: *Expectations and variances.*

1.4 Random variables without a mean

The sum or integral in the definition of the mean, in equations (1.3) and (1.4), might not converge; if it does not, we say that the **mean does not exist**.

For example, let X be a random variable with probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

A random variable with this p.d.f. is said to have a **Cauchy distribution**. It can be checked that f really is a probability density function.

If we attempt to calculate the mean of X, we look at

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx,$$

which should be interpreted as

$$\lim_{s\to\infty, t\to\infty} \int_{-s}^{t} \frac{x}{\pi(1+x^2)}\,dx.$$

However

$$\int_{-s}^{t} \frac{x}{\pi(1+x^2)}\,dx = \frac{1}{2}\left(\log(1+t^2) - \log(1+s^2)\right),$$

and this does not have a well-defined limit as both s and t go to infinity. Hence the mean is undefined.

The Cauchy distribution is not the only example of a distribution without a defined finite mean; there are many others. See, for example, **Q1.10** and **Q1.11**.

The Weak Law of Large Numbers (from MAS113) states that if we have a sequence of independent random variables $X_1, X_2, X_3, \ldots$ with the same distribution and with mean μ, then for any $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}\left[|\bar{X}_n - \mu| > \epsilon\right] \to 0.$$

This tells us that when it exists, we can think of the mean as the long term average of samples of X. However, for a distribution without a defined mean, this result no longer makes sense, because we have no μ.

In fact, we will show later on in this course, that if $X_1, X_2, X_3, \ldots, X_n$ are independent random variables with a Cauchy distribution, $\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$ also has a Cauchy distribution, *regardless* of the value of n. Therefore, there is no (deterministic) value that the sample mean becomes close to, for large n.

Chapter 2

Standard Distributions

Our eventual goal, in this course, is to build statistical models and use them to perform inference; in order to do so we require a library of distributions, to use as building blocks in our models. In this section, we put together such a library.

You will already have met several standard distributions in MAS113. In fact, each 'distribution' is really a family of distributions sharing a common formula for the p.f. or p.d.f. which contains one or more **parameter(s)**.

For example, the binomial family $Bi(n, p)$ has two parameters, n, the number of trials, and p, the success probability. It is common to simply refer to the whole family $Bi(n, p)$ as 'the binomial distribution', and similarly for other (families of) distributions.

The distributions that we choose to include in our library are important for diverse reasons, often

- because they arise from simple models (e.g. the binomial distribution from Bernoulli trials)
- or because they have special mathematical properties (e.g. the normal distribution from the central limit theorem).

Two handouts will be made available, one with a list of standard distributions for discrete random variables, and another with a list of standard continuous distributions.

2.1 Standard discrete distributions

You will already have met many of the most important discrete distributions in first-year courses:

- The **Bernoulli distribution**, written $Bernoulli(p)$, with the single parameter $p \in [0, 1]$, defined by $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$.
- The **binomial distribution**, written $Bi(n, p)$, with two parameters, $n \in \mathbb{N}$ and $p \in [0, 1]$, defined by $\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$, for $k \in \{1, 2, \ldots, n\}$.
- The **Poisson distribution**, written $Poi(\lambda)$ with the single parameter $\lambda \in (0, \infty)$, defined by $\mathbb{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$, for $k \in \mathbb{N}$.
- The **geometric distribution**, written $Geom(p)$, with the single parameter $p \in (0, 1]$, defined by $\mathbb{P}[X = k] = p^k(1 - p)$, for $k \in \mathbb{N}$.

We also introduce two more discrete distributions, which are closely related to the binomial and geometric distributions. Recall that the binomial and geometric distributions both have interpretations in terms of Bernoulli trials. Consider a sequence $(X_i)_{i=1}^{\infty}$ of independent Bernoulli trials, each with success probability p.

- The geometric distribution $Geom(p)$ is the number of trials that we must carry out until we first see success.
- The binomial distribution $Bin(n, p)$ is the number of successes that we will see in the first n trials.

2.1.1 The negative binomial distribution

The **negative binomial distribution** has two parameters, $k \in \mathbb{N}$ and $p \in (0, 1]$. We write it as $NegBin(k, p)$. It is the distribution of the number of (independent) $Bernoulli(p)$ trials we must carry out until we see k successes.

We can use this definition to work out a formula for the probability function of $X \sim NegBin(k, p)$. We can't have k successes before we've done k trials, so $\mathbb{P}[X = r] = 0$ for $r < k$. For $r \in \{k, k+1, k+2, \ldots\}$, we can calculate

$$\begin{aligned}
\mathbb{P}[X = r] &= \mathbb{P}\left[k-1 \text{ successes in first } r-1 \text{ trials, and } r^{th} \text{ trial is a success}\right] \\
&= \mathbb{P}\left[k-1 \text{ successes in first } r-1 \text{ trials}\right] \mathbb{P}\left[r^{th} \text{ trial is a success}\right] \\
&= \binom{r-1}{k-1} p^{k-1}(1-p)^{r-1-(k-1)} \times p \\
&= \binom{r-1}{k-1} p^{k}(1-p)^{r-k}.
\end{aligned}$$

Note that here we use the probability function of the binomial distribution to calculate the probability of seeing $k-1$ successes in the first $n-1$ trials.

The negative binomial distribution is commonly used in sampling, see **Q2.7**.

2.1.2 The hypergeometric distribution

The **hypergeometric distribution** has three parameters, $N \in \mathbb{N}$, $k \in \{0, \ldots N\}$ and $n \in \{0, \ldots, N\}$. If we have a population of N objects, precisely k of which have a special property, and we take a random sample of precisely n objects, then $HypGeom(N, k, n)$ is the number of objects in our sample that has the special property.

Again, we can use this definition to derive a formula for the probability function of $X \sim HypGeom(N, k, n)$. To do so, note that since there are k special objects in our population, $\mathbb{P}[X = r] = 0$ unless $r \in \{0, 1, \ldots, k\}$. For $r \in \{0, 1, \ldots k\}$, we have

$$\begin{aligned}
\mathbb{P}[X = r] &= \frac{\text{number of possible samples of size } n \text{ containing } r \text{ special objects}}{\text{total number of possible samples}} \\
&= \frac{\binom{n}{r}\binom{N-n}{k-r}}{\binom{N}{n}}.
\end{aligned}$$

To see how the denominator is obtained, recall that $\binom{N}{n}$ is the number of ways we can choose n objects from a set of N objects. For the numerator, we must choose r special objects within our sample of size n, and then choose the remaining $k-r$ special objects to be within the $N-n$ objects not included in our sample.

The hypergeometric distribution, as we might expect, is frequently used in combinatorics. Combinatorics is the branch of mathematics that focuses on counting the number of objects that occur in given situations (for example, like the number of different graphs with a vertices and b edges).

2.2 Standard continuous distributions

We now move on to look at some important continuous distributions. Again, you have come some examples of these in MAS113.

We usually define a continuous distribution by writing down its probability distribution function. When we do so, we have to make sure we specify the region on which the p.d.f. is non-zero.

- The **exponential distribution**, written $Exp(\lambda)$, with one parameter $\lambda > 0$, defined by its p.d.f.
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$
- The **uniform distribution**, written $Unif[a, b]$, with two parameters $a < b$, defined by its p.d.f.
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$
It is sometimes convenient to use $Unif(a, b)$, with p.d.f. defined instead to be non-zero on (a, b). The distribution is the same – in both cases the probability of taking the value a or b is zero.

In Sections 2.2.1 and 2.3 we look at some more examples of continuous distributions. Then, in Section 2.4 we will look at using software packages to sketch probability density functions for us.

2.2.1 The (univariate) normal distribution

Again, you will have encountered the normal distribution in MAS113.

We say that X has a **normal distribution** with mean μ and variance σ^2, if the probability density function of X is
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
for all $x \in \mathbb{R}$. We write $X \sim N(\mu, \sigma^2)$. It can be shown (see MAS113) that the mean and variance of a random variable with this p.d.f. really are μ and σ^2.

The special case $N(0,1)$, with p.d.f.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \tag{2.1}$$

is referred to as the **standard normal distribution**. The p.d.f. $\phi(x)$ cannot be integrated explicitly.

Example 6: *Calculating $\mathbb{E}[e^Y]$ where $Y \sim N(0,1)$.*

An important property of the normal distribution family is that if $X \sim N(\mu, \sigma^2)$ and a and b are constants, then

$$aX + b \sim N(a\mu + b, a^2\sigma^2). \tag{2.2}$$

In particular X can be **standardised** by letting $Z = \frac{X-\mu}{\sigma}$, so that $Z \sim N(0,1)$.

Another important property is that, if we have n independent normal random variables $X_1, X_2, \ldots X_n$, with $X_i \sim N(\mu_i, \sigma_i^2)$ then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right). \tag{2.3}$$

If the X_i are not independent, this formula typically does not hold (for example, take $n = 2$ and set $X_2 = -X_1$).

We will prove (2.2) in Example 9 and (2.3) in Section 6.4.

2.2.2 The log-normal distribution

The distribution of $Y = e^X$, where $X \sim N(\mu, \sigma^2)$, is known as the **log-normal distribution**, written $logN(\mu, \sigma^2)$. In Example 11 we will show that the probability density function of $Y \sim logN(\mu, \sigma^2)$ is

$$f_Y(y) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) & \text{if } y \in (0, \infty) \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

The log-normal distribution is useful as a model for a diverse range of quantites, as we will see in Example 35 (which is about particle sizes in aerosols).

2.3 The gamma and beta distributions

In this section we introduce two widely used continuous distributions, namely the gamma and beta families. Many well known distributions, such as the exponential distribution, chi squared distribution, and the uniform distribution, are special cases of these families.

2.3.1 The gamma and beta functions

Recall that the p.d.f. of the standard normal distribution is $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Probability density functions must integrate to give 1, so we have $\int_{-\infty}^{\infty} \phi(x)\, dx = 1$. This means that $\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$, and we can think of the factor $\frac{1}{\sqrt{2\pi}}$ in (2.1) as being placed there 'to make sure ϕ integrates to 1'. We refer to such a factor as a **normalizing constant**

The gamma and beta functions, which we introduce in this section, appear as normalizing constants in the probability density functions of many standard continuous distributions. Before we study these distributions, we need to work out a few facts about the gamma and beta functions.

The **gamma function**, $\Gamma : (0, \infty) \to \mathbb{R}$ is defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du.$$

The first property of the gamma function that we are interested in is that, in a sense, it generalizes the factorial function.

Lemma 2.1 *It holds that* $\Gamma(1) = 1$*, and for* $\alpha > 1$ *we have*

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1).$$

For $n = 1, 2, \ldots$*, we have* $\Gamma(n) = (n-1)!$

Proof: Note that

$$\Gamma(1) = \int_0^\infty e^{-u}\, du = \big[-e^{-u} \big]_0^\infty = 1.$$

Using integration by parts, for any $\alpha > 1$,

$$\begin{aligned}\int_0^\infty u^{\alpha-1} e^{-u}\, du &= \Big[u^{\alpha-1}(-e^{-u}) \Big]_0^\infty - \int_0^\infty (\alpha-1) u^{\alpha-1}(-e^{-u})\, du \\ &= 0 + (\alpha - 1) \int_0^\infty u^{\alpha-2} e^{-u}\, du\end{aligned}$$

That is, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. By repeatedly applying this formula, for any $n \in \mathbb{N}$ we have $\Gamma(n) = (n-1)(n-2)\ldots(3)(2)\Gamma(1)$, and since $\Gamma(1) = 1$ this gives $\Gamma(n) = (n-1)!$ ■

One other specific value, which appears in some formulae for standard distributions, is $\Gamma(1/2) = \sqrt{\pi}$. See **Q2.11**.

In a similar style, the **beta function** is defined for $\alpha, \beta > 0$ by

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du.$$

The beta and gamma functions are related by the formula

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{2.5}$$

The proof of this formula uses a change of variables inside a double integral, and is outside the scope of our course.

Frequently, in the next few sections, we will encounter integrals of the form $\int_0^\infty u^{\alpha-1} e^{-\beta u}\, du$. Note that this integral is similar to the integral defining the Gamma function above, but has an extra constant β. It can be related to the Gamma function by the following change of variables.

Lemma 2.2 *If* $\alpha, \beta > 0$*, we have*

$$\int_0^\infty u^{\alpha-1} e^{-\beta u}\, du = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

PROOF: We apply a change of variables $t = \beta u$. With this substitution, we have

$$\int_0^\infty u^{\alpha-1}e^{-\beta u}\,du = \int_0^\infty \left(\frac{t}{\beta}\right)^{\alpha-1} e^{-t}\left(\frac{1}{\beta}\right)\,dt \;= \beta^{-\alpha}\int_0^\infty t^{\alpha-1}e^{-t}\,dt \;= \beta^{-\alpha}\Gamma(\alpha)$$

as required. ■

2.3.2 The gamma distribution

Let

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & \text{for } x > 0, \\ 0, & \text{for } x \leq 0. \end{cases}$$

The distribution with this p.d.f. is called the **Gamma distribution**. It has two parameters, $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$. If a random variable X has this p.d.f. then we write $X \sim Ga(\alpha, \beta)$.

Let us check that f really is a probability density function. We need to check that $f(x) \geq 0$ for all x, which is clearly true, and that $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Using Lemma 2.2 we have

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\,dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-\beta x}\,dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{\beta^\alpha} \\ &= 1.\end{aligned}$$

So, f really is a p.d.f.

Example 7: *Mean and variance of the Gamma distribution.*

Like the normal distribution, the gamma distribution has several nice properties. In **Q5.6** we will show that if $X_1 \sim Ga(\alpha_1, \beta)$ and $X_2 \sim Ga(\alpha_2, \beta)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim Ga(\alpha_1 + \alpha_2, \beta)$.

The exponential distribution is a special case of the gamma distribution; in fact the $Ga(1, \beta)$ distribution is equal to the $Exp(\beta)$ distribution. To see this, set $\alpha = 1$ in the formula for the p.d.f. and note that it then becomes the p.d.f. of $Exp(\beta)$.

Combining the results of the last two paragraphs, we have that the sum of n independent $Exp(\lambda)$ variables has the $Ga(n, \lambda)$ distribution.

Remark 2.3 *There are alternative parametrisations of the gamma distribution (e.g. $\theta = 1/\beta$), so you sometimes have to be careful with software when using it.*

2.3.3 The chi-squared distribution

The chi-squared[1] distribution, which is used in many statistical tests, is also a special case of the gamma distribution.

[1] Pronounced 'kiy squared'

Let $n \in \mathbb{N}$. If we set $\alpha = n/2$ and $\beta = 1/2$ in the gamma distribution, then we obtain the **chi-squared distribution** with n degrees of freedom, written χ_n. It has a single parameter, $n \in \mathbb{N}$. From the p.d.f. for the gamma distribution, we obtain

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} \exp\left(-\frac{x}{2}\right) & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases}$$

If X has this probability density function then we write $X \sim \chi^2_\nu$. An important special case is when $\nu = 1$, because χ^2_1 is the distribution of Z^2, where $Z \sim N(0,1)$. We will prove this fact in Example 12.

More generally, the χ^2_n is the distribution of the sum of the squares of n independent standard normal random variables; that is if $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim N(0,1)$ and $Y = \sum_{i=1}^n X_i^2$, then $Y \sim \chi^2_n$. See **Q5.7**.

It is this relationship to the normal distribution that makes the ξ^2 distribution important in statistical testing; in many situations statistical errors are (independent and, approximately) normally distributed, and the squares of such errors may therefore be approximated by a chi-squared distribution. We will come back to this idea later on in the course, in Section 5.1.

2.3.4 The beta distribution

Let

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Again, we will show (below) that f is a p.d.f. The distribution with this p.d.f. is called the **Beta distribution** with parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$, written $Be(\alpha,\beta)$.

To see that $f(x)$ is a p.d.f., we note that $f(x) \geq 0$ and

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^1 \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\, dx = 1.$$

We can find the mean and variance of the Beta distribution using similar methods as for the Gamma distribution.

Example 8: *Calculate the mean and variance of X if $X \sim Be(\alpha,\beta)$*

If we take $\alpha = 1$ and $\beta = 1$, we see that the $Be(1,1)$ distribution is the same as the Uniform distribution on $[0,1]$.

The Beta distribution is useful for modelling random quantities which are naturally constrained to be in $[0,1]$ (or, via suitable scaling, in any bounded interval).

2.4 Plotting distributions in R

The aim of this section is to show how to use the computer package R to plot density and distribution functions of random variables.

Most of you will have seen R in use in Level 1 courses. For a more detailed introduction to R, see the course website for the handout "An Introduction to R".

To start with, we assume that R has been installed. We wish to plot the p.d.f. $f_X(x)$ of the random variable $X \sim N(0,1)$. The command we use here is `curve`, which creates a curve of a given function. The form of this command is

```
> curve(f(x), from="lower limit", to="upper limit")
```

or just

```
> curve(f(x), "lower limit", "upper limit")
```

This tells R to plot a curve of a given function $y = f(x)$, where x takes values from "lower limit" to "upper limit". If we don't enter these two limits R will use its own default values. More details on the arguments of the above command can be found by typing `help(curve)`.

Using the command `curve` we can do

```
> curve(dnorm,-3,3)
```

Similarly, one can produce plots of the p.d.f. of any normal variable, $X \sim N(\mu, \sigma^2)$, by using the command `dnorm(x,mean,sd)`. For example

```
> curve(dnorm(x,2,10),-10,14)
```

gives a plot of the p.d.f. of a $N(2, 100)$ variable. Note that `sd` refers to standard deviation, not variance.

Similar plots can be obtained by for the p.d.f.s of other distributions. For the normal distribution use `dnorm`, for the chi-square use `dchisq`, for the Student t use `dt`, for the gamma distribution use `dgamma` (but check the definition of the p.d.f., because there are different ways of paremetrising this distribution), for the beta distribution use `dbeta`, for the binomial use `dbinom`. See also **Q2.6**.

To find the syntax for other distributions, it can be useful to use R's help system, which can be accessed with `help(topic)`.

If the distribution we wish to plot does not exist by default in R, then we can define it in R and plot it using the `curve` command as above. For more information on this, consult the on-line manuals of R.

Chapter 3

Transformations of Continuous Random Variables

The general question here is: if we have a continuous random variable X with a known distribution, and we have another random variable Y defined as

$$Y = g(X)$$

for some function g, then what is the distribution of Y?

For example, in Section 2.3.3 we stated that if X is standard normal and $Y = X^2$ then Y has χ^2_1 distribution. Another example is that if $X \sim N(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

Example 9: *Transforming probabilities 'by hand', standardization of the normal distribution.*

It is more efficient to have a general method of transforming the p.d.f. of X into the p.d.f. of Y. Then, we can use f_Y to calculate probabilities for Y.

Let us look at the case where g is **strictly monotone**, i.e. either strictly increasing or decreasing on the relevant range of x. Note that in this case g has an inverse function g^{-1} which is also strictly increasing (if g is strictly increasing), or strictly decreasing (if g is strictly decreasing).

Recall that R_X denotes the set of values that X may actually take, and that $g(R_X) = \{g(x)\,;\, x \in R_X\}$.

Lemma 3.1 *Suppose that $g : R_X \to \mathbb{R}$ is strictly monotone. Then the p.d.f. of $Y = g(X)$ is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{for } y \in g(R_X), \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

PROOF: If $g^{-1}(y)$ is not in R_X then $f(g^{-1}(y)) = 0$, so it is enough to prove the lemma in the case $R_X = \mathbb{R}$.

Firstly, if g is strictly increasing, then g^{-1} is strictly increasing. Therefore,

$$F_Y(y) = \mathbb{P}[Y \le y] = \mathbb{P}\left[g^{-1}(Y) \le g^{-1}(y)\right] = \mathbb{P}\left[X \le g^{-1}(y)\right] = F_X(g^{-1}(y)).$$

Differentiating with respect to y, using the chain rule we have

$$f_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) = f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y). \tag{3.2}$$

Secondly, if g is strictly decreasing, then g^{-1} is strictly decreasing. So,

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}\left[g^{-1}(Y) \geq g^{-1}(y)\right] = \mathbb{P}\left[X \geq g^{-1}(y)\right] = 1 - F_X(g^{-1}(y)).$$

Note that here, because g^{-1} is decreasing, we must change the sign when we apply g^{-1} to both sides of an inequality. Differentiating with respect to y,

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y). \tag{3.3}$$

If g^{-1} is increasing then $\frac{d}{dy} g^{-1}(y) \geq 0$, and if g^{-1} is decreasing, $\frac{d}{dy} g^{-1}(y) \leq 0$. So, we can combine our two cases from (3.2) and (3.3), into (3.1). ■

To apply this lemma, we must check all its conditions, which takes several steps. Usually this means we must:

1. write down the p.d.f. of X and identify the set R_X on which $f_X(x) > 0$,
2. write down a function g such that $Y = g(X)$,
3. check that g is strictly monotone and find g^{-1}, $\frac{dg^{-1}}{dy}$, and $g(R_X)$.

Having done so, we can apply the lemma and deduce $f_Y(y)$ from (3.1). This is our usual method for carrying out univariate transformations.

Example 10: *The cube root of the $Be(3,1)$ distribution.*

There are many examples of this technique in the exercises, several of which display relationships between standard distributions. One such relationship is transforming the normal distribution into the log-normal distribution, as in the next example.

Example 11: *The log-normal distribution.*

In general, if one of the above steps fails, for example if g is not strictly monotone, then we work 'by hand'; we have to identify the problem and find a way around it.

This happens (for example) in one of transformations we mentioned in Chapter 2, namely taking the square of a normal random variable to obtain a chi-squared random variable. In this case we have $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = x^2$. Of course, this g is not strictly monotone.

Example 12: *Square of a standard normal (the χ^2_1 distribution).*

Sometimes, operating directly with the p.d.f. is not possible. For example, if $g : \mathbb{R} \to \mathbb{R}$ is given by $g(x) \equiv 0$ then, regardless of X, Y is a discrete random variable and $\mathbb{P}[Y = 0] = 1$. In this case, Y does not have a probability density function.

Chapter 4

Multivariate Distribution Theory

Often, we are interested in several random quantities at once, and these quantities may affect each other. For example, if we were interested in a possible link between traffic and pollution in a city center we might record, on a given day:

X_1 = number of cars travelling into the city center

X_2 = an indicator of air quality (e.g. parts per million of carbon monoxide)

X_3 = wind speed

X_4 = rainfall.

As is often the case in statistics, we'd need to record this information for multiple days, before we would expect to discover a link.

In such cases, we have a **multivariate** random variable or **random vector**

$$\mathbf{X} = (X_1, X_2, \ldots, X_k),$$

where $k \in \mathbb{N}$ is the number of different types of observation. Formally, $\mathbf{X}$ is a mapping from the sample space S into k-dimensional space $\mathbb{R}^k$. Of particular interest will be how the components $X_1, X_2, \ldots, X_k$ vary together.

We need to upgrade our definitions of continuous/discrete random variables to the multivariate case.

Definition 4.1 *We say that* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ *is a*

- ***continuous random vector****, if all* $X_1, X_2 \ldots, X_k$ *are continuous random variables.*
- ***discrete random vector****, if all* $X_1, X_2 \ldots, X_k$ *are discrete random variables.*

In any other case we say that $\mathbf{X}$ *is neither continuous, nor discrete.*

You will already have met discrete random vectors in MAS113. In this chapter we will discuss continuous random vectors; later we will introduce the important case of vectors with multivariate normal distributions and also look at transformations of random vectors.

The case $k = 1$ is the univariate case, from Chapter 1. In this chapter we will concentrate on the case $k = 2$, which is known as the **bivariate** case, but the same ideas extend to general $k \in \mathbb{N}$. We will tend to write $X = X_1$ and $Y = X_2$, giving $\mathbf{X} = (X, Y)$.

4.1 Joint distribution and density functions

Our first step is to discuss how distribution functions and probability density functions apply to random vectors.

Definition 4.2 *The **joint distribution function** of the random vector* $\mathbf{X} = (X, Y)$ *is the function* $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ *given by*

$$F_{X,Y}(x, y) = \mathbb{P}[X \leq x, Y \leq y]$$

In principle, as in the univariate case, if we know the value of this function for all x, y then we can calculate any probability involving X and Y.

Definition 4.3 *If the partial derivative*

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = f_{X,Y}(x, y)$$

exists then we say that $f_{X,Y}$ *is the **joint probability density function (p.d.f.)** of* X *and* Y.

The joint probability density function is analogous to the p.d.f. of a single random variable. To find the probability that the pair (X, Y) lies in some region D of the plane then we must integrate $f_{X,Y}$ over D; in other words

$$\mathbb{P}[(X, Y) \in D] = \iint_D f_{X,Y}(x, y)\, dx\, dy. \tag{4.1}$$

Pictorially, if we plot the surface $z = f_{X,Y}(x, y)$ in three dimensions then $\mathbb{P}[(X, Y) \in D]$ is equal to the volume between the surface $z = f_{X,Y}(x, y)$ and the image of D in the plane $z = 0$. Note that if we choose $D = (-\infty, x] \times (-\infty, y]$ in (4.1) we get

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv.$$

Since evaluating probabilities involves double integration, often over a bounded region, we must take care to get the limits of integration right. We'll discuss this in the examples below (and also in the solutions to **Q4.1** and **Q4.3**).

In the univariate setting, we saw cases where different formulas for the p.d.f. were needed on different intervals. We allow similar cases in the bivariate (and multivariate!) situations too; the boundaries separating regions in which different formulas for the p.d.f. apply may now be lines as well as points.

Example 13: *Joint probability density functions.*

We can use the joint p.d.f. to evaluate expectations of $g(X, Y)$, where $g : \mathbb{R}^2 \to \mathbb{R}$. To do so, we use the formula

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy. \tag{4.2}$$

4.2 Marginal and conditional distributions

When we have a random vector (X, Y), we can also think of both X and Y as (univariate) random variables.

Definition 4.4 *Given a random vector (X, Y), with joint p.d.f. $f_{X,Y}(x, y)$ the **marginal p.d.f.** of X is $f_X : \mathbb{R} \to \mathbb{R}$ given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}\, dx.$$

In words, we integrate out the y component. To find the marginal distribution of Y, we integrate out the x component.

When viewed as a univariate random variable, f_X is the p.d.f. of X; we say 'marginal' p.d.f. so as we don't forget that X is also part of the random vector (X, Y). The marginal distribution of X is the distribution with p.d.f. $f_X(x)$.

Example 14: *Marginal distributions.*

We can also think of the distribution of X *given* that Y takes a particular value, say y. We could view this as taking the bivariate pair (X, Y) and artificially imposing the condition that $Y = y$, then asking what the distribution of X will be. This captures the fact that value of Y might affect the value of X.

Definition 4.5 *The **conditional p.d.f.** of X given $Y = y$ is $f_{X|Y=y} : \mathbb{R} \to \mathbb{R}$ given by*

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \tag{4.3}$$

which is defined only when $f_Y(y) > 0$. If we swap the roles of X and Y, we obtain $f_{Y|X=x}(y)$.

It can be shown that, provided $f_Y(y) > 0$, the formula (4.3) does genuinely define a p.d.f. (as a function of x), see Q**4.11**. It is also common to write $f_{X|Y}(x|y)$ for the conditional p.d.f., although we will always write $f_{X|Y=y}(x)$. The conditional distribution of X given $Y = y$, is the distribution of a random variable with p.d.f. $y \mapsto f_{X|Y=y}(x)$.

Example 15: *Conditional distributions.*

4.3 Independence, covariance and correlation

Recall that, by definition, a pair (X, Y) of random variables are **independent** if and only if

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\, \mathbb{P}[Y \in B]$$

for all $A, B \subseteq \mathbb{R}$. If we have independent random variables X and Y with probability density functions $f_X(x)$ and $f_Y(y)$ respectively, then it is easy to find the joint p.d.f. of the random vector (X, Y):

Lemma 4.6 *A pair of random variables X and Y are independent if and only if*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all x, y.

PROOF: Suppose X and Y are independent. Then

$$F_{X,Y}(x,y) = \mathbb{P}[X \le x, Y \le y] = \mathbb{P}[X \le x]\mathbb{P}[Y \le y] = F_X(x)F_Y(x).$$

Differentiating both sides of this equation with respect to both x and y gives $\frac{\partial F_{X,Y}}{\partial x \partial y} = \frac{\partial F_X}{\partial x}\frac{\partial F_Y}{\partial y}$, which by definition of the (joint) marginal p.d.f. means that $f_{X,Y}(x,y) = f_X(x)f_Y(x)$.

Alternatively, suppose that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Then, for any $A, B \subseteq \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{P}[X \in A, Y \in B] &= \int_A \int_B f_X(x)f_Y(y)\,dy\,dx \\
&= \int_A f_X(x) \left(\int_B f_Y(y)\,dy \right) dx \\
&= \mathbb{P}[Y \in B] \int_A f_X(x)\,dx \\
&= \mathbb{P}[Y \in B]\mathbb{P}[X \in A].
\end{aligned}$$

So X and Y are independent. ■

Alternatively, if we are given a p.d.f. $f_{X,Y}(x,y)$, there is a simple test for independence. It has the extra advantage that it does require us to find f_X or f_Y explicitly.

Corollary 4.7 *A pair of random variables X and Y are independent if and only if*

$$f_{X,Y}(x,y) = g(x)h(y) \tag{4.4}$$

for some pair of functions $g(x)$ and $h(y)$.

PROOF: Firstly, if X and Y are independent then Lemma 4.6 shows that (4.4) holds, just take $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.

Alternatively, if we know (4.4) holds then

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy = g(x) \int_{\mathbb{R}} h(y)\,dy$$

and similarly $f_Y(y) = h(y)\int_{\mathbb{R}} g(x)\,dx$. Using (4.4) a second time gives

$$1 = \iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy = \iint_{\mathbb{R}^2} g(x)h(y)\,dx\,dy = \left(\int_{\mathbb{R}} g(x)\,dx \right) \left(\int_{\mathbb{R}} h(y)\,dy \right)$$

and hence

$$\begin{aligned}
f_{X,Y}(x,y) &= \left(\iint_{\mathbb{R}^2} g(x)h(y)\,dx\,dy \right)^{-1} \times \left(g(x) \int_{\mathbb{R}} h(y)\,dy \right) \times \left(h(y) \int_{\mathbb{R}} g(x)\,dx \right) \\
&= 1 \times f_X(x) \times f_Y(y).
\end{aligned}$$

So, by Lemma 4.6, X and Y are independent. ■

Usually, when we want to show independence we use Corollary 4.7, because then all we need to do is factorize (and we don't mind if g and h are p.d.f.s or not). If we are lucky, we can then recognize $g(x)$ and $h(y)$ as p.d.f.s, and then we can write both X and Y as standard distributions.

Example 16: *Independence and factorizing $f_{X,Y}$.*

When X and Y are not independent, we can try to measure how much they depend on each other. Let us write $\mu_X = E[X]$ and $\mu_Y = E[Y]$, along with $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$.

Definition 4.8 *The **covariance** of X and Y is defined as*

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Definition 4.9 *The **correlation coefficient** of X and Y is defined as*

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{(\mathrm{Var}(X)\,\mathrm{Var}(Y))}}.$$

Here, $\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ and $\mathbb{E}[XY]$ can be calculated using (4.2), for example to calculate $\mathbb{E}[XY]$ use $g(x,y) = xy$ to obtain

$$\mathbb{E}[XY] = \int\int_{\mathbb{R}^2} xy f_{X,Y}(x,y)\, dy\, dx.$$

A useful fact is that, if X and Y are independent, then we have

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0. \tag{4.5}$$

However, it is easy to find examples of random variables with zero correlation that are *not* independent, see **Q4.6**.

Correlation and covariance measure the extent to which X and Y vary together; if both X and Y tend to have the same sign then $\mathrm{Cov}(X,Y)$ will be positive, but if X and Y tend to have different signs, they are negative.

Example 17: *Covariance and correlation.*

If you have a set of data which are a random sample from the distribution of (X,Y), then 'Pearson's sample correlation coefficient', which some of you will have seen, is an estimator of $\rho(X,Y)$.

4.4 Conditional expectation

The **conditional expectation** of X given that Y takes the value y is defined as the expectation of a random variable with the conditional distribution:

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\, dx.$$

Note that the formula above is a function of y; we could write it as $g(y) = \mathbb{E}[X|Y = y]$. We define

$$\mathbb{E}[X|Y] = g(Y),$$

which formally means that $\mathbb{E}[X|Y]$ is a random variable and, for each element s of the sample space, $\mathbb{E}[X|Y](s) = g(Y(s)) = \mathbb{E}[X|Y = Y(s)]$.

Example 18: *Calculating conditional expectation.*

We define the conditional variance as the variance of the conditional distribution of X given Y. In symbols,

$$\mathrm{Var}(X|Y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 \,|\, Y\right].$$

We can also define conditional covariances: if X, Y and Z are random variables, then the conditional covariance of X and Y, given Z, is

$$\mathrm{Cov}(X, Y|Z) = \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z].$$

Conditional variances and covariances are closely related to (unconditioned) variances and covariances, by the following lemma.

Lemma 4.10 *It holds that*

1. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$.
2. $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$.
3. $\mathrm{Cov}(X, Y) = \mathbb{E}[\mathrm{Cov}(X, Y \,|\, Z)] + \mathrm{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z])$.

The proof of property 1 is:

Example 19: *Proof of Property 1.*

The proof of properties *2* and *3* are similar in style, and we don't include them in this course.

Properties *1-3* are useful when the best way of finding the mean and variance of (say) X is by conditioning on Y, or if we already know the distribution of X given Y.

Example 20: *Calculation of expectation and variance by conditioning.*

Chapter 5

Transformations of Multivariate Distributions

In Chapter 3 we looked at univariate transformations, and in particular how to find the p.d.f. of $g(X)$ from the p.d.f. of X. Recall that, in the one dimensional case, if $U = g(X)$ (for monotone g) then we had

$$f_U(u) = f_X\left(g^{-1}(u)\right) \times \left|\frac{dg^{-1}}{du}\right|. \tag{5.1}$$

for u within the image of the range of X, that is $u \in g(R_X)$.

We now ask this same question for a multivariate random variable $\mathbf{X} = (X_1, \ldots, X_k)$. We will concentrate here on the bivariate case $k = 2$, but the theory described can be extended to the general case.

Suppose that we have two continuous random variables, X and Y, with joint p.d.f. $f_{X,Y}(x, y)$. We have a transformation $u = u(x, y)$ and $v = v(x, y)$, which we use to define two new random variables, $U = u(X, Y)$ and $V = v(X, Y)$. We are interested to find the joint p.d.f. of (U, V).

We require that the transformation used be 'genuinely two dimensional', in the sense that the whole transformation is continuous, differentiable and one-to-one. Consequently, it is possible to find an inverse of the transformation: $x = x(u, v)$ and $y = y(u, v)$.

In the bivariate case, the equivalent to (5.1), is the formula

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) \times \left|\det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}\right|, \tag{5.2}$$

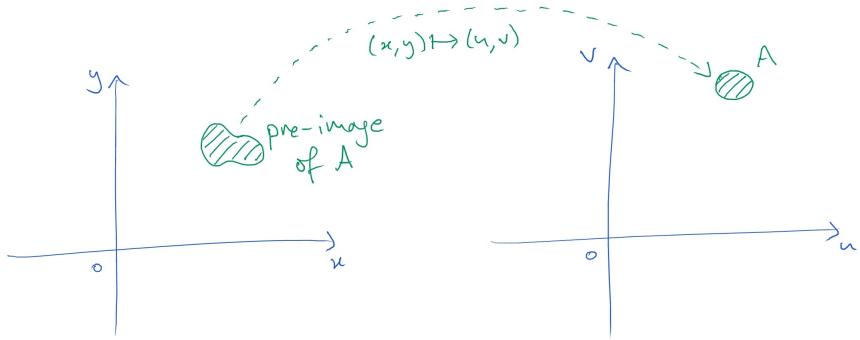
which is valid for all (u, v) in the image of the range of (X, Y), and $f_{U,V}(u, v)$ is zero otherwise.

Remark 5.1 *Comparing* (5.2) *to* (5.1)*, we see that they are very similar. They both involve writing our initial p.d.f. in terms of the our variables, and multiplying by a term involving derivatives of the inverse transformation.*

Let us briefly explain where the factor

$$J = \left|\det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}\right|$$

comes from. We call J the **Jacobian**[1]. Recall that $f_{X,Y}(x,y)$ is a measure of how likely (X,Y) is to be (infinitesimally) close to (x,y). We want $f_{U,V}(u,v)$ to be a measure of how likely (U,V) is to be close to (u,v). Let A be a small region containing (u,v). To find out which values of (x,y) would fall into this region, we need to find out how big the pre-image of A is in the (x,y) plane.



It turns out, that the factor by which the area of A changes, in the above diagram, is precisely $|\det(J)|$. See MAS211 for details of this; it is really just a change of variables $(x,y) \mapsto (u,v)$ for the double integral $\iint f_{X,Y}(x,y)\,dx\,dy$.

It is important to identify the range of values taken by (U,V), and it usually helps to draw a sketch of the transformation. In particular, if X and Y take values in a restricted range given by inequalities in x and y, then these must be translated into inequalities in u and v by substituting for x and y in terms of u and v.

To summarise the steps, we must:

1. Write down a transformation $u = u(x,y)$ and $v = v(x,y)$ such that $U = u(X,Y)$ and $V = v(X,Y)$.
2. Find the inverse transformation $x = x(u,v)$ and $y = y(u,v)$, and calculate $|\det(J)|$.
3. Find region of (u,v) that corresponds to region of (x,y) for which $f_{X,Y}(x,y) > 0$. Usually a sketch is very helpful.

Having done so, we can the apply the formula

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}\big(x(u,v), y(u,v)\big) \times |J| & \text{if } f_{X,Y}\big(x(u,v), y(u,v)\big) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This process is, in many ways, best learned through examples.

Example 21: *Transforming bivariate random variables.*

Example 22: *The Box-Muller transform, simulating normal random variables.*

Sometimes we are interested in only one transformed random variable, $U = g(X,Y)$ say. In this case one possibility is to simply choose a V (but we must still make sure the transformation

[1] Or, more precisely, the determinant of the Jacobian matrix.

$(x, y) \to (u, v)$ is genuinely two-dimensional). We can then find $f_{U,V}(u, v)$ as above, and integrate out v to obtain $f_V(v)$.

If there is no obvious choice for V, using $V = X$ or $V = Y$ often works well.

Example 23: *Summing Gamma random variables.*

5.1 Sample mean, sample variance and Student's t distribution

In this section we add one one distribution to our library of distributions; namely **Student's t distribution**, which is used in many statistical tests. Using multivariate transformations, we can explain why Student's t distribution becomes important in statistics.

Student's t distribution is a continuous distribution with a single parameter, $n \in \mathbb{N}$, and probability density function given by

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \tag{5.3}$$

for all $x \in \mathbb{R}$. We write $X \sim t_n$. This distribution has a close connection to sample mean and sample variance, which is explored in a series of questions on the exercise sheets. We summarize the results of these questions here.

Suppose that we have a data set $x_1, \ldots, x_n$, each of which is a real number. In many situations, we can model $(x_1, \ldots, x_n)$ as a sequence of (samples from) independent normal random variables, with unknown mean and variance. With this in mind, we consider a sequence $(X_1, \ldots, X_n)$ of i.i.d. $N(\mu, \sigma^2)$ random variables.

The sample mean and variance of $(X_1, \ldots, X_n)$ are, respectively,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

In **Q5.7**, we show that the χ^2_n chi-squared distribution is the sum of the squares of n i.i.d. standard normals. Using this fact, in **Q6.8** we show that

$$\frac{(n-1)s^2}{\sigma^2} \text{ has a } \chi^2_{n-1} \text{ distribution.}$$

On the other hand, the sample mean is a sum of independent normal distributions. Using (2.3), this means that the sample mean $\bar{X}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. Therefore, using standardization (2.2),

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \text{ has a } N(0, 1) \text{ distribution.}$$

Let Z be a $N(0, 1)$ random variable and let W be a chi-squared random variable with n degrees of freedom, where Z and W are independent. In **Q5.8**, we show that

$$X = \frac{Z}{\sqrt{W}/\sqrt{n}}$$

has the t distribution and n degrees of freedom. Therefore, with $n-1$ in place of n, using what we know about the distributions of the sample mean and variance, the statistic

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{s} = \frac{\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)}{\sqrt{\frac{(n-1)s^2}{\sigma^2}} \Big/ \sqrt{n-1}} \tag{5.4}$$

has Student's t distribution with $n-1$ degrees of freedom (this is **Q6.9**).

The point of all this is: we now know that $T \sim t_{n-1}$, and in particular *the distribution of T does not depend on μ or σ^2*. At first glance this is very surprising, because equation (5.4) does contain μ and σ^2, as well as the (X_i). Crucially, this special property of T means, even if we don't know μ and σ^2, we can still use T in statistical tests.

Chapter 6

The Multivariate Normal Distribution

The multi-variate normal is the most important continuous joint distribution, and is commonly used to model multivariate data. Besides being a natural model in many situations, the multivariate normal has several nice properties that make it easy to work with.

Before we begin to look at the multivariate normal in detail, we need to set up two further pieces of theory. For the duration of this chapter we write a^T for the transpose of (the vector or matrix) a.

6.1 Covariance matrices, mean vectors and linear transformations

In this section, we look at a general random vector, $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$.

Definition 6.1 *The **mean vector** of* $\mathbf{X}$ *is the vector*

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_k] \end{pmatrix}.$$

We will often write $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$, where $\mu_i = \mathbb{E}[X_i]$. It is often more convenient to write column vectors, in which case we write $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_k])^T = (\mu_1, \ldots, \mu_k)^T$.

Definition 6.2 *Then **covariance matrix** of* $\mathbf{X}$*, denoted by* $\mathrm{Cov}(\mathbf{X})$*, is the* $k \times k$ *matrix in which the* $(i,j)^{th}$ *element is* $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$*. That is,*

$$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \ldots & \mathrm{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_k, X_1) & \mathrm{Cov}(X_k, X_2) & \ldots & \mathrm{Cov}(X_k, X_k). \end{pmatrix}$$

Since $\text{Var}(X_i) = \text{Cov}(X_i, X_i)$, this matrix has the variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$ of the X_i along its leading diagonal. Note that have $\sigma_{ii} = \sigma_i^2$. From the definition of correlation coefficient we may also write $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ where ρ_{ij} is the correlation coefficient between X_i and X_j.

The covariance matrix is a symmetric matrix, because $\sigma_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = \sigma_{ji}$. If the $X_1, X_2, \ldots X_k$ are independent (or merely uncorrelated) then Σ is a diagonal matrix, meaning that only the diagonal elements σ_{ii} are non-zero.

Example 24: *Mean vectors and covariance matrices.*

Matrix notation is useful when we consider linear transformations of $\mathbf{X}$. Let $\mathbf{A}$ be a $m \times k$ matrix and $\mathbf{b}$ be a vector with m components (both non-random). We denote a linear transformation of $\mathbf{X}$ as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

so that $\mathbf{Y}$ is a vector with m components.

Lemma 6.3 *It holds that*

$$\begin{aligned}\mathbb{E}[\mathbf{Y}] &= \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b} \\ \text{Cov}(\mathbf{Y}) &= \mathbf{A}\,\text{Cov}(\mathbf{X})\mathbf{A}^T\end{aligned}$$

PROOF: The proof relies on linear algebra and it is outside the scope of this course; we include it for completeness. Since multiplying by a matrix is a linear operation, and $\mathbb{E}$ is linear (i.e. it satisfies $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$), we get

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{A}\mathbf{X} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b}.$$

For the second part, we note that $\text{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right]$ where the expectation is taken componentwise. Note that here, $(\mathbf{X} - \boldsymbol{\mu})$ is a $1 \times k$ matrix and $(\mathbf{X} - \boldsymbol{\mu})^T$ is a $k \times 1$ matrix; they are multiplied together as matrices to give a $k \times k$ matrix. Since $\mathbb{E}[\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b}$, writing $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ we have

$$\begin{aligned}\text{Cov}(\mathbf{Y}) &= \mathbb{E}\left[(\mathbf{Y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})(\mathbf{Y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^T\right] \\ &= \mathbb{E}\left[\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{A}(\mathbf{X} - \boldsymbol{\mu}))^T\right] \\ &= \mathbb{E}\left[\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{A}^T\right] \\ &= \mathbf{A}\mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right]\mathbf{A}^T \\ &= \mathbf{A}\,\text{Cov}(\mathbf{X})\mathbf{A}^T\end{aligned}$$

as required. ■

Example 25: *Linear transformation of a random vector.*

Example 26: *Special case: Variance of a sum.*

6.2 The bivariate normal distribution

Consider two independent random variables U and V, each with the (univariate) standard normal distribution; that is $U \sim N(0,1)$ and $V \sim N(0,1)$, both with pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

By independence (Lemma 4.6), the joint p.d.f. of U and V is given by the product of the individual p.d.f.s,

$$f_{U,V}(u,v) = f_U(u) f_V(v) = \frac{1}{2\pi} e^{-(u^2+v^2)/2} \tag{6.1}$$

for $(u,v) \in \mathbb{R}^2$. This is a first example of our multivariate normal.

More generally, we are interested in the case of a pair of (non-standard) normal random variables, $N(\mu_1, \sigma_1^2)$ and $N(\mu_1, \sigma_2^2)$, and we are interested in the case where they are not, necessarily, independent.

Definition 6.4 *The **bivariate normal distribution** $\mathbf{X} = (X_1, X_2)^T$, with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and covariance matrix $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{smallmatrix}\right)$ is the distribution with p.d.f.*

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \exp\left(-\frac{\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \right)$$

We write $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Example 27: *The case where X_1 and X_2 are independent.*

Example 28: *Plotting the p.d.f. of the bivariate normal.*

It will take a little time to explain where the formula above comes from; it is the p.d.f. that results from a linear transformation of a pair of independent standard normals. We'll use a combination of Chapter 5 and Lemma 6.3 to see this.

Let

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

be a non-singular 2×2 matrix, and let $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ be a 2-vector. We now consider the random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{S} \begin{pmatrix} U \\ V \end{pmatrix} + \boldsymbol{\mu}.$$

Here, (U, V) are a pair of independent standard normals.

Let's think first about the mean and covariance matrix of $(X_1, X_2)^T$. The transformation S is linear, so we calculate these using Lemma 6.3. They are

$$\mathbb{E}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = S\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \mathrm{Cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{S}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{S}^T = \mathbf{S}\mathbf{S}^T \tag{6.2}$$

$$= \begin{pmatrix} s_{11}^2 + s_{12}^2 & s_{22}s_{12} + s_{21}s_{11} \\ s_{22}s_{12} + s_{21}s_{11} & s_{21}^2 + s_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}. \tag{6.3}$$

The last line here is the definition of σ_{ij}, in terms of the s_{ij}. Therefore, σ_1^2 and σ_2^2 are the variances of X_1 and X_2, and σ_{12} is their covariance.

Now, let's move on to think about the joint p.d.f. of $(X_1, X_2)^T$. Since

$$\begin{aligned} X_1 &= s_{11}U + s_{12}V + \mu_1, \\ X_2 &= s_{21}U + s_{22}V + \mu_2, \end{aligned} \tag{6.4}$$

we can view $(X_1, X_2)^T$ as a bivariate transformation of $(U, V)^T$. So, we can use the method from Section 5 to transform the probability density function. We know the p.d.f. of $(U, V)^T$ from (6.1).

The forward transformation is given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{S}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

and the inverse transformation is given by

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{S}^{-1}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} = \frac{1}{\det \mathbf{S}}\begin{pmatrix} s_{22} & -s_{21} \\ -s_{12} & s_{11} \end{pmatrix}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$$

Therefore,

$$u = \frac{1}{\det \mathbf{S}}(s_{22}(x_1 - \mu_1) - s_{12}(x_2 - \mu_2)), \quad v = \frac{1}{\det \mathbf{S}}(-s_{21}(x_1 - \mu_1) + s_{11}(x_2 - \mu_2)),$$

and (after a short calculation) the Jacobian of the inverse transformation is $|1/\det S|$. Hence, the joint p.d.f. $f_{X_1,X_2}(x_1, x_2)$ of X_1 and X_2 is

$$\frac{1}{2\pi|\det \mathbf{S}|}\exp\left\{-\frac{\left[(s_{22}(x_1 - \mu_1) - s_{12}(x_2 - \mu_2))^2 + (-s_{21}(x_1 - \mu_1) + s_{11}(x_2 - \mu_2))^2\right]}{2(\det \mathbf{S})^2}\right\},$$

which, with a little work, can be rearranged as

$$\frac{1}{2\pi|\det S|}\exp\left\{-\frac{\left[\sigma_2^2(x_1 - \mu_1)^2 + \sigma_1^2(x_2 - \mu_2)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2)\right]}{2(\det \mathbf{S})^2}\right\}. \tag{6.5}$$

Finally, by (6.2) we have $\det \Sigma = \det(\mathbf{S}\mathbf{S}^T) = (\det \mathbf{S})^2$, so we can replace $|\det S|$ by $\sqrt{\det \Sigma}$ in the above. As $\det \Sigma = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$, we can re-write the joint p.d.f. f_{X_1,X_2} as

$$\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}}\exp\left\{-\frac{\sigma_2^2(x_1 - \mu_1)^2 - 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) + \sigma_1^2(x_2 - \mu_2)^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)}\right\},$$

for all $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$. This matches the p.d.f. in Definition 6.4.

Remark 6.5 *For any positive definite covariance matrix Σ, it is possible to find a non-singular matrix S such that (6.2) holds. A positive definite $k \times k$ matrix M is a matrix for which $a^T M a > 0$ for all non-zero vectors $a \in \mathbb{R}^k$.*

To prove this claim we must use some linear algebra; this part is outside the scope of our course but we include it here for completeness. Any covariance matrix (because it is symmetric) can be diagonalised as $\Sigma = PDP^{-1} = PDP^T$, where P is orthogonal and D is diagonal. Moreover, D has positive entries on its diagonal (because Σ is positive definite). So D can be written as $\hat{D}^2$, and hence $\Sigma = P\hat{D}\hat{D}P^T$. If we let $S = P\hat{D}Q$ for any orthogonal matrix Q then, using orthogonality,

$$SS^T = P\hat{D}Q(P\hat{D}Q)^T = P\hat{D}QQ^T\hat{D}P^T = P\hat{D}\hat{D}P^T = \Sigma.$$

So any positive definite Σ can be obtained from some S.

6.2.1 Higher dimensions

The joint p.d.f. of the bivariate normal can also be written in terms of $\boldsymbol{\mu}$ and Σ as

$$f(X_1, X_2) = \frac{1}{2\pi(\det(\Sigma))^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

This p.d.f. is well defined for any symmetric positive definite 2×2 matrix Σ.In this form, we could also take $\boldsymbol{\mu}$ to be a vector in $\mathbb{R}^k$, and Σ to be a $k \times k$ matrix. Doing so gives the p.d.f. of the general **multivariate normal distribution** $N_k(\boldsymbol{\mu}, \Sigma)$.

6.3 Marginal distributions, correlation and covariance

For the remainder of Chapter 6, let $\mathbf{X} = (X_1, X_2)^T$ be a bivariate normal with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{smallmatrix}\right)$.

We now go on to investigate several properties of the bivariate normal distribution. All the properties we study can be generalize to the multi-variate normal, in any dimension.

Lemma 6.6 *The marginal distributions of $\mathbf{X}$ are $X_1 \sim N(\mu_1, \sigma_{11})$ and $X_2 \sim N(\mu_2, \sigma_{22})$.*

PROOF: One way to see this is to 'integrate out', using the method of Section 4.2 with the bivariate normal p.d.f., but there is an easier way.

During derivation of the bivariate normal in section 6.2. In (6.4) we wrote write the components of a bivariate normal random vector $\mathbf{X}$ as a linear combination of independent normals and constants. The theory of linear combinations of univariate normal distribution, from (2.3), tells us that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. ■

We can also ask precisely when the components of the bivariate normal are independent. It turns out that this can be checked using a very simple condition.

Lemma 6.7 *The two components, X_1 and X_2, of a bivariate normal distribution are independent if and only if* $\mathrm{Cov}(X_1, X_2) = 0$.

PROOF: If $\text{Cov}(X_1, X_2) \neq 0$, then X_1 and X_2 are not independent. On the other hand, if $\text{Cov}(X_1, X_2) = \sigma_{12} = \sigma_{21} = 0$, then we can factorise the p.d.f. of $(X_1, X_2)^T$ (from Definition 6.4) as follows:

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{\sigma_2^2(x_1-\mu_1)^2 + \sigma_1^2(x_2-\mu_2)^2}{2(\sigma_1^2\sigma_2^2)}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_1-\mu_1)^2}{\sigma_1}\right) \times \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{(x_1-\mu_2)^2}{\sigma_2}\right). \end{aligned}$$

We can recognize this as the product of two (univariate) normal p.d.f.s, which are independent by Lemma 4.6. ■

It is important to remember that this result does not hold for general random variables; in general $\text{Cov}(X, Y) = 0$ does not imply independence! But it does hold for components of multivariate normals.

Example 29: *Marginal distributions of the bivariate normal, and their covariance.*

6.4 Linear transformations of the bivariate normal

We obtained the bivariate normal distribution by applying a linear transformation to a pair of independent standard normals. As a result, it is natural to ask what happens if we apply a linear transformation to a bivariate normal.

Let $\mathbf{A}$ be a 2×2 matrix and let $\mathbf{b}$ be a 2×1 vector. We define

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b},$$

so as $\mathbf{Y}$ is a linear transformation of $\mathbf{X}$.

Lemma 6.8 *Suppose that* $\mathbf{A}$ *is non-singular. Then the random vector* $\mathbf{Y}$ *has a bi-variate normal distribution, with mean vector* $\mathbf{A}\boldsymbol{\mu}$ *and covariance matrix* $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.

PROOF: In Section 6.2 (in particular, in Remark 6.5), we showed that

$$\mathbf{X} = \mathbf{S}\mathbf{U} + \boldsymbol{\mu}$$

for some matrix $\mathbf{S}$, where $\mathbf{U} = (U, V)^T$ is a pair of independent (univariate) standard normals. So we can write

$$\mathbf{Y} = \mathbf{A}\mathbf{S}\mathbf{U} + \mathbf{A}\boldsymbol{\mu} + \mathbf{b} = (\mathbf{A}\mathbf{S})\mathbf{U} + (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}).$$

Both $\mathbf{A}$ and $\mathbf{S}$ are non-singular, so $\mathbf{A}\mathbf{S}$ is also non-singular. Hence, $\mathbf{Y}$ is a linear transformation of $\mathbf{U}$ and we can apply our theory from Section 6.2 (with $\mathbf{A}\mathbf{S}$ in place of $\mathbf{S}$ and $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ in place of $\mathbf{b}$). It follows immediately that $\mathbf{Y}$ has a bivariate normal distribution with mean vector and covariance matrix

$$\begin{aligned} \mathbb{E}[\mathbf{Y}] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\ \text{Cov}[\mathbf{Y}] &= \mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^T = \mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T, \end{aligned}$$

as required. ■

Example 30: *Transformations of the bivariate normal*

One special case of this, is the case where $\mathbf{A}$ is a row vector (i.e. a 1×2 matrix!) and $\mathbf{b} = b$ by a scalar (a 1×1 vector). This gives $Y = A\mathbf{X} + b$ as a scalar. Lemma 6.8 tells us that Y has a normal distribution, and that $\mathbb{E}[Y] = \mathbf{A}\boldsymbol{\mu} + b$, $\mathrm{Var}(Y) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$, both of which are scalars.

Using this special case, and choosing $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mathbf{b} = 0$ shows that, for independent $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ we have

$$\mathbf{Y} = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1 + \sigma_2).$$

Iterating this equation provides the (long delayed!) proof of (2.3).

6.5 Conditional distributions are normal

After seeing, in Section 6.3, that the marginals distributions of the bivariate are themselves (univariate) normals, it should not be surprising to learn that the conditional distributions of bivariate normal distributions are also normal.

Let us write ρ for the correlation of X_1 and X_2, that is $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$.

Lemma 6.9 *The conditional distribution of X_2 given that $X_1 = x_1$ is a normal distribution with mean $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ and variance $(1 - \rho^2)\sigma_2^2$.*

Proof: We could prove this by dividing the joint p.d.f. (which we know from Definition 6.4) by the marginal p.d.f. (which we know from Lemma 6.6) to obtain the conditional probability density function of X_2 given that $X_1 = x_1$. Instead, we will use a simpler argument that involves Lemma 6.8.

Let $Y_1 = X_1$ and $Y_2 = X_2 - \lambda X_1$, for some $\lambda \in \mathbb{R}$ which we will choose later. Write $\mathbf{Y} = (Y_1, Y_2)^T$. Then $\mathbf{Y} = \mathbf{A}\mathbf{X}$ where $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix}$. By Lemma 6.8, we have

$$\mathbb{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 - \lambda\mu_1 \end{pmatrix}$$

and

$$\begin{aligned} \mathrm{Cov}(\mathbf{Y}) &= \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & -\lambda\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 \\ -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 & \lambda^2\sigma_1^2 - 2\lambda\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix}. \end{aligned}$$

We now choose $\lambda = \rho\sigma_1^{-1}\sigma_2$, which means that the non-diagonal terms of the above matrix are zero, giving

$$\mathrm{Cov}(\mathbf{Y}) = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & (1 - \rho^2)\sigma_2^2 \end{pmatrix}.$$

This means that $\text{Cov}(Y_1, Y_2) = 0$, and using Lemma 6.7, we have that Y_1 and Y_2 are independent. Since $Y_1 = X_1$, this means that X_1 and Y_2 are independent. We have

$$X_2 = \lambda X_1 + Y_2.$$

If we condition on $X_1 = x_1$, then Y_2 is unchanged since it is independent of X_1. Hence, conditional on $X_1 = x_1$ we have $X_2 = \lambda x_1 + Y_2$, so we have that X_2 is normally distributed with mean $\lambda x_1 + \mathbb{E}[Y_2] = \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) + \mu_2$ and variance $\text{Var}(Y_2) = (1 - \rho^2)\sigma_2^2$. ■

Example 31: *Conditional distributions for bivariate normal*

6.6 Higher dimensions

Suppose that we are dealing with $\mathbf{X} = (X_1, \ldots, X_k)$, a multivariate normal for general $k \in \mathbb{N}$. We refer to k as the dimension of the multivariate normal. The results of the previous sections (which were stated with $k = 2$) extend naturally to general $k \in \mathbb{N}$, and let us briefly mention these extensions.

In particular, if we pick any subset of the $X_1, \ldots, X_k$, their marginal distribution will still be a (now, possibly multivariate) normal. Applying any non-singular linear transformation results in another multivariate normal, of the same dimension. A pair of components, X_i and X_j are independent if and only if $\text{Cov}(X_i, X_j) = 0$. Moreover, if we condition on the values of any subset of the $X_1, \ldots, X_k$, the distribution of the other components will be a multivariate normal (with parameters that can be calculated explicitly, but we will not give a formula here).

Recall that the univariate normal p.d.f. cannot be integrated explicitly, with the consequence that (in general) normal probabilities have to be approximated numerically. High dimensional numerical integration is a hard problem, even for a powerful computer, and multivariate normal probabilities can be very difficult to evaluate.

Example 32: *Linear transformation of a three dimensional normal distribution.*

Chapter 7

Likelihood

We will now look at **statistical inference**, which means analysing data to obtain information about the processes which produced the data. In particular, we will be looking at methods of inference based on likelihood. Many common methods of data analysis rely on likelihood.

7.1 Likelihood

In Chapter 2, we built up a library of standard distributions. Each of these distributions had one or more parameters; for example the Poisson distribution $Poi(\theta)$ has the single parameter $\theta > 0$. Up to this point, we viewed the parameters as constants; the key idea of likelihood is to view the parameters as variables.

7.1.1 Recap: maximising functions

We will need to find the global maximums of functions, as a key part of our inference methods. You should already know how to do this, but let us briefly recap.

Suppose that $I \subseteq \mathbb{R}$ and that we have a differentiable function $f : I \to \mathbb{R}$. We say that a point x_0 **maximises** f if

$$f(x_0) \geq f(x) \text{ for all } x \in I.$$

That is, if x_0 is the location of the global maximum value of $f(\cdot)$.

Two key facts:

$$\text{A point } x_0 \in I \text{ is a turning point of } f \text{ if and only if } \left.\frac{df}{dx}\right|_{x=x_0} = 0. \tag{7.1}$$

$$\text{A turning point } x_0 \text{ of } f \text{ is a local maximum if } \left.\frac{d^2f}{dx^2}\right|_{x=x_0} < 0. \tag{7.2}$$

One useful note is that, if a differentiable function f has a single turning point, and this turning point is a local maximum, then it is automatically the global maximum.

Example 33: *Maximisation of a function*

If more than one turning point appears, we have to be more careful (and, in this course, we will approach such cases through curve sketching or by using `R`).

7.1.2 Discussion

Let us first illustrate the idea with an example. Suppose that we know (or suspect, or hope!) that it is sensible to use the Geometric distribution $Geom(\theta)$ to model the number of times we have to roll a biased die before we get a 6. What we don't know, is which value of θ is best to use. Clearly, we must have $\theta \in [0, 1]$ because we use the Geometric distribution, but how should we choose exactly which value of θ to use?

Since we now care about the value of the parameter(s) θ, we will tend to write probability density functions as $f(x; \theta)$, and probability functions as $p(x; \theta)$. In this case, the Geometric distribution has probability function $p(x; \theta) = \theta^x(1 - \theta)$.

Suppose, for now, that we have a just single item of data; we roll the die 5 times until we first see a 6. We will worry about handling multiple data points later. If we let $X \sim Geom(\theta)$, the probability of this event is

$$p(5; \theta) = \theta^5(1 - \theta).$$

This is a function of θ.

Now, here is the key idea: *since we observed the value* 5*, it would make sense if we chose* θ *to make* $p(5; \theta) = \mathbb{P}[X = 5]$ *as large as possible.* That is, we want to choose θ so as our model $Poi(\theta)$ is as likely as possible to reproduce the data that we actually observed. So, what it comes down to, is finding the value of θ which maximises the function

$$L(\theta; 5) = p(5; \theta) = \theta^5(1 - \theta)$$

amongst the range of possible choices of θ, in this case $\Theta = (0, \infty)$. We call $L(\theta; 5)$ the likelihood of the parameter value θ, given the data 5. We write the (hopefully, unique) maximiser of $L(\theta; 5)$ as $\hat{\theta}$, and we call it the maximum likelihood estimator of θ. We can find $\hat{\theta}$ using the method from Section 7.1, in fact from Example 33 we know that $\hat{\theta} = \frac{5}{6}$.

7.1.3 Maximum Likelihood Estimation I

In what situations can we use the method from Section 7.1.2?

We might want to use a continuous random variable in our model. Then we use the probability density function instead of the probability function. The theory in this case is very similar, so it is common to use the same notation in both cases. We will do so, only in this chapter! Therefore, from now on

$$\textit{in discrete examples we will have } f_X(x) = \mathbb{P}[X = x].$$

We will discuss more general situations, including how to use multiple data points, in Section 7.2. For now, we are ready to define likelihood.

To summarise, let X be a random variable with a known distribution, with p.d.f. (or p.f.) $f(x; \theta)$, where θ is an unknown parameter. Let Θ be the set of possible values of θ. Let x be our data point, which we view as a sample of X.

Definition 7.1 *The* ***likelihood function*** *of X, given the data x, is $L : \Theta \to \mathbb{R}$ defined by*

$$L(\theta; x) = f(x; \theta)$$

*We refer to the value of $L(\theta; x)$ as the **likelihood of θ given the data** x.*

*The (hopefully unique!) $\theta \in \Theta$ which maximises $L(\theta; x)$ is known as the **maximum likelihood estimator** of θ. We usually denote it by $\widehat{\theta}$.*

Sometimes, we will refer to the likelihood function of a distribution; this is the likelihood function of a random variable with that distribution. The process of finding the maximum likelihood estimator $\hat{\theta}$ is known as maximum likelihood estimation.

Note that when we view the likelihood function as a function of θ, it is not a probability density function. For example, it typically will not integrate over θ to give 1.

Example 34: *Likelihood functions and maximum likelihood estimators*

The value $\hat{\theta}$ which maximises $L(\theta; x)$ changes if use a different value for x. This is natural - the choice of parameters that we think is best, depends on the data that we have.

7.2 Models and data

In Section 7.1.3 we estimated a parameter value based on a single data point. Using only one data point is highly unreliable, and in this section we discuss how to carry out maximum likelihood estimation using many data points.

We begin by defining some common terminology used in statistical inference.

7.2.1 Data

Typically we will have a set of n **data points** which we can think of as a vector, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We will think of the data as being realisations of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Note the use of capital letters for the random variables and lower case letters for the values they take.

The random vector $\mathbf{X}$ will have a joint p.d.f.

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \ldots, x_n).$$

This p.d.f. will be unknown, the aim of the inference being to obtain information about it.

We will assume that our data points $x_1, x_2, \ldots, x_n$ come from independent, identically distributed experiments. With this in mind, we call them **i.i.d. samples**. Because of this, we also assume that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed. In this case, the joint p.d.f. $f_{\mathbf{X}}(\mathbf{x})$ will be a product of terms for each experiment:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f(x_i), \tag{7.3}$$

where $f(x)$ is the common p.d.f. of the random variables $X_1, X_2, \ldots, X_n$.

7.2.2 Models and Parameters

We assume that we already know that the joint p.d.f. of $\mathbf{X}$ takes a particular form, usually involving some standard distribution. We refer to this as our **model**. However, the **parameters**

of this standard distribution are unknown, and our aim in analysing the data will be to obtain good choices of values for these parameters, based on the data we have.

Remark 7.2 *It may seem odd to declare that f is unknown, and then assume that in fact f takes a particular form with only unknown parameters. There are statistical methods aimed at handling completely unknown f, but they are outside of the scope of this course. In many situations it is sensible to assume a carefully chosen model with unknown parameters.*

Our choice of model may well be wrong. But we hope that it is approximately correct, and in future statistics course you will discover that there are so-called 'goodness of fit' tests to help us check.

We denote the parameters of our model by $\boldsymbol{\theta}$; we represent $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ as a vector, although in any particular case the set of parameters can be a scalar (a single number), a matrix or some other structure. We write Θ for the set of possible parameter values.

Sometimes some of the unknown parameters are so-called **nuisance parameters**: their values are unknown, and we have to take account of this in our analysis, but they are not what we are really interested in.

Given a model, and a particular set of parameter values $\boldsymbol{\theta}$, we write the p.d.f. of $\mathbf{X}$ as $f(\mathbf{x}; \boldsymbol{\theta})$, to make sure that we don't forget the importance of the parameters. As before, we use the same notation in the discrete case, but it now means a probability function.

Example 35: *Models, parameters and data (aerosols).*

7.2.3 Maximum Likelihood Estimation II

How can we apply the ideas of maximum likelihood estimation in our new setting? The key is to note that Definition 7.1 already makes sense in the multivariate case, with our model $\mathbf{X}$ in place of X and a vector of data $\mathbf{x}$ in place of x. We also allow more than just one unknown parameter.

So, to summarise, let $\mathbf{X} = (X_1, \ldots, X_k)$ be a random variable, with a known distribution that has one or more unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_j)$. Write Θ for the set of all possible choices $\boldsymbol{\theta}$ of parameter(s). Let $\mathbf{x}$ be our vector of data, which we think of as a sample of $\mathbf{X}$.

Definition 7.3 *The **likelihood function** of $\mathbf{X}$, given the data $\mathbf{x}$, is the function $L : \Theta \to \mathbb{R}$ defined by*

$$L(\boldsymbol{\theta}; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}).$$

*The value $\boldsymbol{\theta} \in \Theta$ which maximises $L(\boldsymbol{\theta}; \mathbf{x})$ is known as the **maximum likelihood estimator** of $\boldsymbol{\theta}$, written $\widehat{\boldsymbol{\theta}}$.*

As usual, here $f(\mathbf{x}; \theta)$ is the probability function if $\mathbf{X}$ is discrete, or the probability density function if $\mathbf{X}$ is continuous.

We are mainly interested in the case where our data are i.i.d. samples, and we assume the $X_1, X_2, \ldots, X_n$ are identically distributed. In this case, from (7.3), the likelihood function of

our model is

$$L(\boldsymbol{\theta}; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}). \tag{7.4}$$

where f is the common p.d.f. of the X_i.

Example 36: *Maximum likelihood estimation with i.i.d. data.*

Example 37: *Maximum likelihood estimation (radioactive decay).*

7.3 Maximisation Techniques

Maximum likelihood estimation comes down to a maximisation problem. Whether this is easy or difficult depends on (a) the statistical model we use in the form $f(\mathbf{x}|\boldsymbol{\theta})$ and (b) the parameter vector $\boldsymbol{\theta}$. One-parameter problems are clearly easier to handle and in many cases multi-parameter problems require the use of numerical maximisation techniques.

7.3.1 Log-likelihood

When maximising $L(\boldsymbol{\theta}; \mathbf{x})$ it is usually easier to work with the logarithm of the likelihood instead of the likelihood itself. In this course we always work with natural logarithms. These work well when dealing with the many standard distributions whose p.d.f.s include an exponential term.

Definition 7.4 *Given a likelihood function $L(\boldsymbol{\theta}; \mathbf{x})$, the* ***log-likelihood function*** *is*

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \log L(\boldsymbol{\theta}; \mathbf{x}).$$

Maximising $\ell(\boldsymbol{\theta}; \mathbf{x})$ over $\boldsymbol{\theta} \in \Theta$ produces the same estimator $\widehat{\boldsymbol{\theta}}$ as maximising $L(\boldsymbol{\theta}; \mathbf{x})$, because the function $\log(\cdot)$ is strictly increasing. However, maximising ℓ is usually easier!

Using the log-likelihood is by far the most important maximisation technique. Part of the reason is that $\log(ab) = \log(a + b)$, so in the case of i.i.d. data points, from (7.4) we have

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \log(L(\boldsymbol{\theta}; \mathbf{x})) = \log\left(\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \log(f(x_i; \boldsymbol{\theta})).$$

Using ℓ instead of L changes the $\prod$ into a $\sum$, and it is usually easier to work with a sum than a product.

Example 38: *Maximum likelihood estimation through log-likelihood (mutations in DNA).*

7.3.2 Discrete parameters

When we maximise $L(\boldsymbol{\theta}; \mathbf{x})$ (or $l(\boldsymbol{\theta}; \mathbf{x})$), we need to be careful to keep $\boldsymbol{\theta}$ within the parameter set Θ. In most of the examples we will meet in this module $\boldsymbol{\theta}$ will be continuous and so we can use differentiation to obtain the maximum. However, in some cases, such as the next example, the possible values of $\boldsymbol{\theta}$ may be discrete (i.e. Θ is a discrete set) and in such cases we cannot use differentiation.

Remark 7.5 *Note that saying $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}, \boldsymbol{x})$ is continuous is not the same thing as saying that the distribution of $\mathbf{X}$ is continuous!*

Example 39: *Maximum likelihood estimation for discrete parameters (mass spectroscopy).*

7.3.3 Multi-parameter problems

For multi-parameter problems, where $\boldsymbol{\theta}$ is a vector, a similar procedure can be followed. Here for simplicity we consider only the case where there are 2 parameters (so that $\boldsymbol{\theta}$ is a 2×1 vector) and write $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Now we find a stationary point $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)$ of the log-likelihood by solving the simultaneous equations

$$\frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x})}{\partial \theta_1} = 0, \quad \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x})}{\partial \theta_2} = 0. \tag{7.5}$$

These equations are the analogue of (7.1); that is, of looking for turning points in the one parameter case by solving $\frac{df}{dx} = 0$. In two dimensions and higher, the turning points that we find may be maxima or minima, or saddle points.

To check that a turning point is a (local) maximum, we have to check this an analogue of equation (7.2). First we calculate the so called **Hessian matrix**:

$$H = \begin{pmatrix} \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial \theta_1^2 & \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial \theta_1 \partial \theta_2 \\ \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial \theta_1 \partial \theta_2 & \partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial \theta_2^2 \end{pmatrix}$$

and then we evaluate H at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is the stationary point we found using (7.5).

In the 2 variable case we can use a fact from multi-variable calculus: if

$$\left.\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1^2}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} < 0 \quad \text{and} \quad \left.\det H\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} > 0 \tag{7.6}$$

then we can conclude that our turning point is a local maximum.

Example 40: *Multi-parameter maximum likelihood estimation (rainfall).*

Remark 7.6 *In the general multivariate case, to check that a turning point is a local maxima we should check that H, when evaluated at the turning point, is a negative definite matrix. This fact is outside of the scope of our course, but we mention it here for completeness.*

A negative definite $k \times k$ matrix $\mathbf{M}$ is a matrix for which $\mathbf{a}^T \mathbf{M} \mathbf{a} < 0$ for all non-zero vectors $\mathbf{a} \in \mathbb{R}^k$. When $k = 2$ this is equivalent to (7.6). For example, you can easily check that $-I$, where I is the identity matrix, is negative definite.

7.3.4 Using a computer

In some cases, particularly when a complex model is used, or when many parameters are unknown, it is not possible to obtain an expression for the maximum likelihood estimator $\hat{\theta}$.

These cases can be approached with the aid of a computer, and **machine optimization**, which means using a computer to try and approximate the maximum value of the likelihood function. There are a wide range of algorithms designed to maximise functions numerically, but this is outside the scope of our current course.

There are also other methods of statistical inference! See later courses.

7.3.5 A warning example

Sometimes, we have to be very careful about using differentiation to maximise the likelihood function. We illustrate with an example.

Example 41: *Maximum likelihood estimation for the uniform distribution*

The moral of the story is: if something seems strange during maximisation, draw a picture of the function you are trying to maximise.

7.4 Quantifying uncertainty

Maximum likelihood estimation gives us a single value for the unknown parameters $\boldsymbol{\theta}$, a so-called point estimate. In many settings in statistical inference we want to go further than point estimation, in particular to give some idea of the uncertainty in our point estimate. For example, where we are trying to estimate a single parameter θ, we may want to produce an interval estimate, typically a set of values $[\theta_1, \theta_2]$ which we believe that the true value θ lies in. Alternatively, we may want to test a hypothesis about θ. The likelihood function can often be used to construct appropriate methods in these settings too, and as with maximum likelihood estimation it can often be shown that they are in some sense optimal.

We will start off by thinking about interval estimation. Assume, in the one parameter case, that we have a likelihood function $L(\theta; \mathbf{x})$ defined for $\theta \in \Theta$, maximised at its maximum likelihood estimate $\hat{\theta}$. Then a natural choice of interval estimate is to set some threshold, L_0 say, and to use the values of θ such that $L(\theta; \mathbf{x}) \geq L_0$ as an interval estimate. One common choice for the threshold is to choose L_0 to be a fixed multiple of the maximum likelihood, say

$$L_0 = e^{-k} L(\hat{\theta}; \mathbf{x})$$

for some chosen $k > 0$. Equivalently in terms of the log-likelihood,

$$\log L_0 = \ell(\hat{\theta}; \mathbf{x}) - k.$$

Our choice of k here will involve a trade off between a precise answer (meaning a narrow interval) and minimising the risk of missing the true value from the interval: a small k will give a narrow interval but relatively low confidence that the interval contains the true value, while a large k will give a larger interval and higher confidence.

More generally, we can make the following definition. The **k-unit likelihood region** for parameters $\boldsymbol{\theta}$ based on data $\mathbf{x}$ is the region

$$R_k = \left\{ \boldsymbol{\theta} : L(\boldsymbol{\theta}; \mathbf{x}) \geq e^{-k} L(\hat{\boldsymbol{\theta}}; \mathbf{x}) \right\},$$

or equivalently

$$R_k = \left\{ \boldsymbol{\theta} : \ell(\boldsymbol{\theta}; \mathbf{x}) \geq \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) - k \right\},$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ based on $\mathbf{x}$.

The values of $\boldsymbol{\theta}$ within the k-unit likelihood region are those whose likelihood is at least within a factor e^{-k} of the maximum. For instance, points in the 1-unit region have likelihoods

within a factor $e^{-1} = 0.368$ of the maximum. The 2-unit region contains points with likelihoods within a factor $e^{-2} = 0.135$ of the maximum. The 2-unit region is the most commonly used in practice.

Example 42: *Likelihood regions*

If we are trying to test a null hypothesis $H_0 : \theta = \theta_0$ against a general alternative hypothesis $H_1 : \theta \neq \theta_0$, then we can use a similar idea: we choose a suitable k, construct the k-likelihood region R_k, and accept H_0 if θ_0 is inside R_k, or reject H_0 if θ is outside R_k.

Example 43: *Hypothesis tests based on likelihood*

This leads into the idea of so-called 'likelihood ratio tests', which you will see more of if you take further courses in statistics.

Chapter 8

Case Studies

Maximum likelihood is one of the most widely used methods of statistical inference. It has many variants and extensions, some of which can be found in future statistics courses. In this chapter, we look at two case studies, taken from the recent literature, in which maximum likelihood estimators were used in 'real world' problems.

Our two case studies are chosen somewhat at random; many many other examples exist and could equally fill their place. Our first case study comes from ecotoxicology, which is the science of measure the impact of toxins on the environment, and focuses on asking if expert opinions can be used as a substitute for experimental data. The second concerns clinical trials; the process in which drugs are analysed to determine if they are fit for general use.

8.1 Ecotoxicology

We look at a study[1] relating to of the levels of toxic chemicals found in rivers.

The study was part of the process of setting standards for toxic chemicals in rivers. The aim is to discover safe levels for the concentrations of pollutants, aiming to protect aquatic animals. However, there is a very large variety of species to protect; fish, snails, insects, leeches, etc. Data on the toxicity of any given chemical is available for only a small number of the species that are of interest.

Due to the shortage of toxicity data, the study was trying to find out if it was possible to use the expertise of freshwater biologists as a substitute for (expensive, lengthy) experiments to produce more toxicity data. We will focus our attention on one particular toxin, the insecticide *chlorpyrifos*.

Our data comes in two parts:

- All the available toxicity data for chlorpyrifos. The data are experimentally obtained estimates, for a small number of species, of the LC50 (the concentration that will kill 50% of the individuals) after a 96-hour exposure to chlorpyrifos.

[1] Grist, E.P.M., O'Hagan, A., Crane, M., Sorokin, N., Sims, I. and Whitehouse, P. (2006). Bayesian and time-independent species sensitivity distributions for risk assessment of chemicals. *Environmental Science and Technology* **40**, 395–401.

Species	*Taxon*	*96hr LC50* ($\mu g/L$)	*Expert* (*average sensitivity score*)
Anguilla anguilla	Anguillidae (eels)	540	4.17
Asellus aquaticus	Asellidae (water hoglice)	2.7	4.08
Caenis horaria	Caenidae (mayflies)	0.5	5.86
Chironomus tentatus	Chironomidae (midges)	0.47	3.56
Corixa punctata	Corixidae (lesser waterboatmen)	2	5.11
Rutilus rutilus	Cyprinidae (carp)	120	4.08
Gammarus lacustris	Gammaridae (shrimps)	0.11	5.57
Pungitus pungitus	Gasterosteidae (sticklebacks)	4.7	4.13
Peltodytes sp.	Haliplidae (water beetles)	0.8	5.00
Leptocerida sp.	Leptoceridae (caddis flies)	0.77	6.00
Oncorhynchus mykiss	Salmonidae (salmon)	7.1	5.40

Figure 8.1: LG50 and expert opinions on the effects of chlorpyrifos.

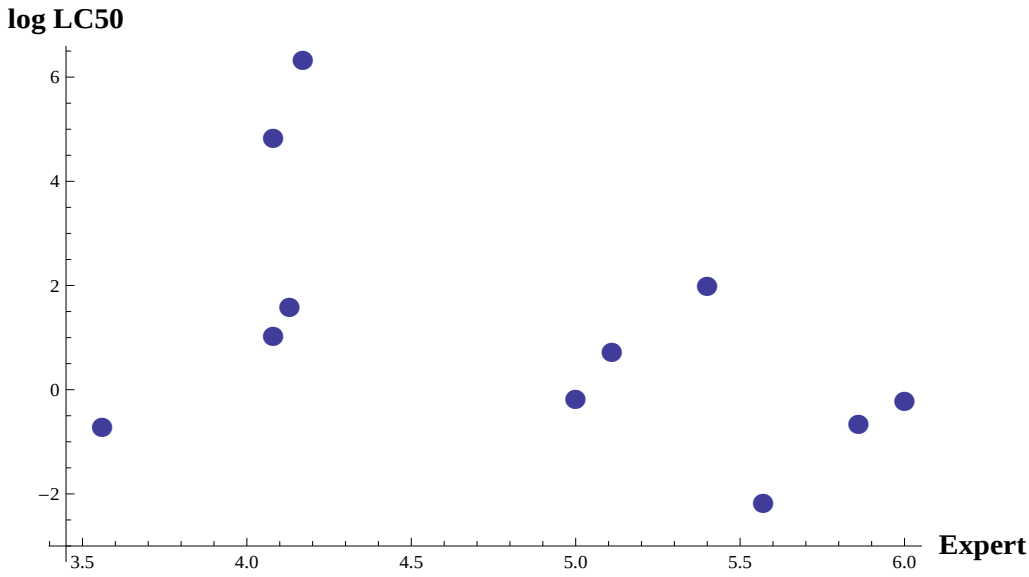


Figure 8.2: The log of LC50 plotted against the expert sensitivity score.

- Estimated scores, from freshwater biologists, on a scale of 1-8, of the sensitivity of these species to chlorpyrifos.

The aim is to work out if we can correlate the experimental data with the estimated scores; if we can then we have grounds to hope that our expert opinions are reliable to be used for a wider range of species (and if so, we might use their opinions to set levels for the species on which we don't have good experimental data).

Our data is tabulated in Figure 8.1. In Figure 8.2, we see a graph of the logs of the LG50 scores, plotted against the experts sensitivity scores. This plot suggests an (approximately) linear relationship between them. As we expect, when experts estimate low sensitivity, a higher concentration is needed to damage the organism. More precisely, lets build a model.

First, we need some notation for the data in Figure 8.1. For $i = 1, 2, \ldots, 11$, let $y_i = \log z_i$, where z_i is the ith toxicity measurement, and let x_i be the corresponding expert sensitivity score.

The following statistical model will be assumed for these data. We suppose that a linear regression relationship applies between these variables, so that

$$y_i = \alpha + \beta x_i + \epsilon_i \ ,$$

where the ϵ_is are independent $N(0, v)$ errors, which would often be called **noise**, and $\boldsymbol{\theta} = (\alpha, \beta, v)$ are the unknown parameters. This type of model is often known as a **linear model**.

Our first step is find estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{v}$ for the parameters, using maximum likelihood. We expect that $\hat{\beta}$ will turn out to be negative, because increasing sensitivity to chlorpyrifos should (intuitively) be associated with a lower LC50; moreover it is suggested by Figure 8.2. We are then interested in comparing the likelihood of $\beta = 0$ (i.e. no correlation) to that of $\beta = \hat{\beta}$.

We have $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, $i = 1, \ldots, 11$, and the observations are assumed independent. So (writing $v = \sigma^2$ again), the likelihood

$$L(\alpha, \beta, v; \mathbf{y}) = \prod_{i=1}^{11} \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2v}\right),$$

and the log likelihood is thus

$$\ell(\alpha, \beta, v; \mathbf{y}) = -\frac{11}{2}(\log(2\pi) + \log v) - \frac{1}{2v}\sum_{i=1}^{11}(y_i - \alpha - \beta x_i)^2.$$

From this, because there is no v in the second term on the right hand side, we can start by minimising the 'sum of squares' term $\sum_{i=1}^{11}(y_i - \alpha - \beta x_i)^2$. After doing so, we can then maximise with respect to v. This type of maximisation will be studied in the second half of MAS223 and there is no need for us to go into details here. After maximisation (with the help of `R`), it turns out that our MLEs are

$$\hat{\alpha} = 7.78, \quad \hat{\beta} = -1.39, \quad \hat{v} = 4.48.$$

The log likelihood at the MLE is -23.86.

If β is assumed to be zero, we get MLEs of 1.10 for α and 5.72 for v. The log likelihood of these values is -25.20. A difference of 1.34 is generally not high, so we do not get strong evidence for $\beta \neq 0$.

8.2 Predicting the outcome of clinical trials

We look at a study[2] concerning clinical trials. Clinical trials typically come in phases I, II and III, which are done in increasing order of size, complexity and (consequently) cost. A successful drug must pass through all three stages.

The study is interested in modelling the results of phases I and II, which are assumed to have already happened, and then using this model to predict the outcome of phase III. These predictions can help to *design* the phase III trial. We will now try and describe how this can be done, using a simplified[3] version of their model.

[2] De Ridder, F. (2005), Predicting the Outcome of Phase III Trials using Phase II Data: A Case Study of Clinical Trial Simulation in Late Stage Drug Development. Basic & Clinical Pharmacology & Toxicology, 96: 235-241

[3] The model also incorporated multiple time steps, an extra factor relating to the variability of the drugs effect on individual patients, and allowed for the possibility of a placebo effect; we omit all of these!

The effect of the drug on the patients was measured using a so-called 'symptom score', which is arrived at by a combination of medical tests, doctors opinions, patient questionnaires, etc. A symptom score R_i is measured for the i^{th} patient at the start of the trial. At the end of the trial, aftera second symptom score S_i is also measured. The model is

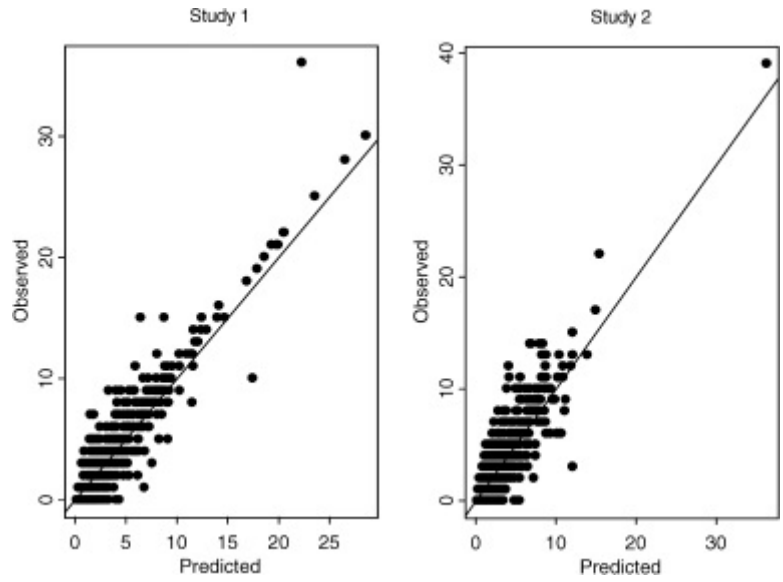
$$S_i = R_i + N_i^{\lambda} + \alpha(d_i)^{\beta}.$$

Here:

- d_i is the dose given to patient i, and $\alpha \in (0, \infty)$ and $\beta \in \mathbb{R}$ are unknown parameters that describe how the effect of the drug varies for different doses.
- N_i^{λ} is an (independent) $Poi(\lambda)$ random variable, with parameter $\lambda \in (0, \infty)$, which models the noise involved in measuring patients symptom scores.

The (already obtained) data from the earlier phases was used to find maximum likelihood estimators $\hat{\alpha}, \hat{\beta}$ and $\hat{\lambda}$. The maximisers were found numerically, using SAS (which is a similar software package to `R`).

Once the MLEs had been found, the real data was compared against the results of simulating data from the model with its parameters set to the MLEs. For two (entirely separate) drugs, the article represents the result of doing this with a graph:



We can see that the model, with its parameters set to $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$, compares favourably with known data. Therefore, we hope that the model could predict the outcome of larger trials with some degree of accuracy. To do so, we pretend that our model *is* reality and use it to simulate data. This predicts how the phase III trial might look, which in turn allows us to improve the design of the (much larger) phase III trial.

For example, for one of the drugs, it was asked: if we carried out a phase III trial in which we gave 1000 patients the drug and compared their results to those of 1000 patients who received a placebo, would we expect to see a statistically significant result? If not, would increasing the sample size help?